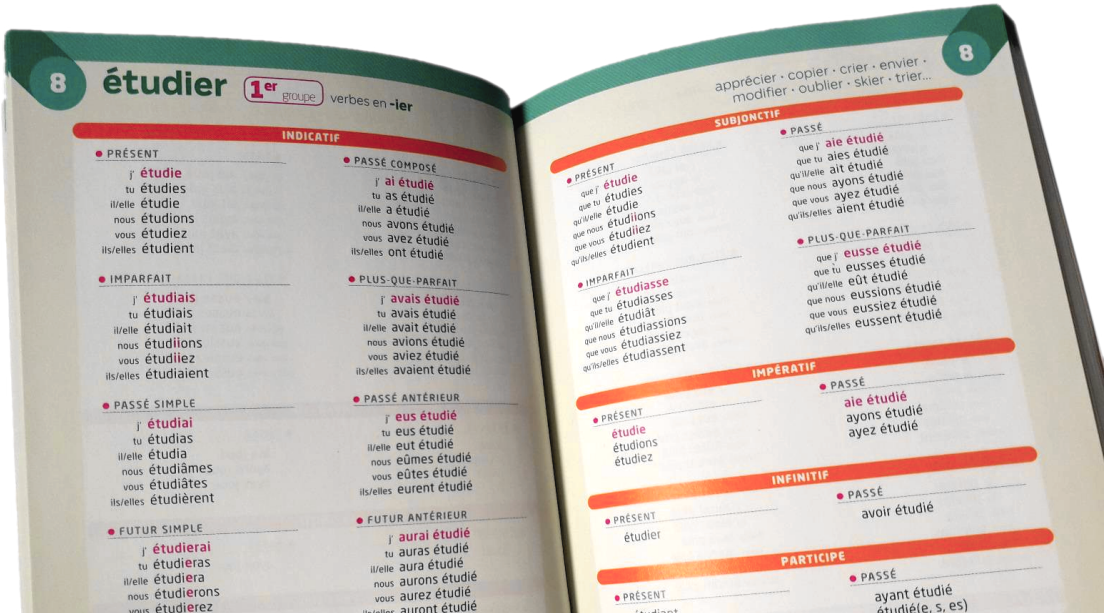
Computational morphology needs better lexical data

Sacha Beniamine Webinaire ILFC, 16th of April 2025

Inflection



8 étudier 1er groupe verbes en -ier

INDICATIF

PRÉSENT
j' étudie
tu étudies
il/elle étudie
nous étudions
vous étudiez
ils/elles étudient

IMPARFAIT
j' étudiais
tu étudiais
il/elle étudiait
nous étudiions
vous étudiiez
ils/elles étudiaient

PASSÉ SIMPLE
j' étudiai
tu étudias
il/elle étudia
nous étudiâmes
vous étudiâtes
ils/elles étudièrent

FUTUR SIMPLE
j' étudierai
tu étudieras
il/elle étudiera
nous étudierons
vous étudierez

PASSÉ COMPOSÉ
j' ai étudié
tu as étudié
il/elle a étudié
nous avons étudié
vous avez étudié
ils/elles ont étudié

PLUS-QUE-PARFAIT
j' avais étudié
tu avais étudié
il/elle avait étudié
nous avions étudié
vous aviez étudié
ils/elles avaient étudié

PASSÉ ANTÉRIEUR
j' eus étudié
tu eus étudié
il/elle eut étudié
nous eûmes étudié
vous eûtes étudié
ils/elles eurent étudié

FUTUR ANTÉRIEUR
j' aurai étudié
tu auras étudié
il/elle aura étudié
nous aurons étudié
vous aurez étudié

8

apprécier · copier · crier · envier · modifier · oublier · skier · trier…

SUBJONCTIF

PRÉSENT
que j' étudie
que tu étudies
qu'il/elle étudie
que nous étudiions
que vous étudiiez
qu'ils/elles étudient

IMPARFAIT
que j' étudiasse
que tu étudiasses
qu'il/elle étudiât
que nous étudiassions
que vous étudiassiez
qu'ils/elles étudiassent

PASSÉ
que j' aie étudié
que tu aies étudié
qu'il/elle ait étudié
que nous ayons étudié
que vous ayez étudié
qu'ils/elles aient étudié

PLUS-QUE-PARFAIT
que j' eusse étudié
que tu eusses étudié
qu'il/elle eût étudié
que nous eussions étudié
que vous eussiez étudié
qu'ils/elles eussent étudié

IMPÉRATIF

PRÉSENT
étudie
étudions
étudiez

PASSÉ
aie étudié
ayons étudié
ayez étudié

INFINITIF

PRÉSENT
étudier

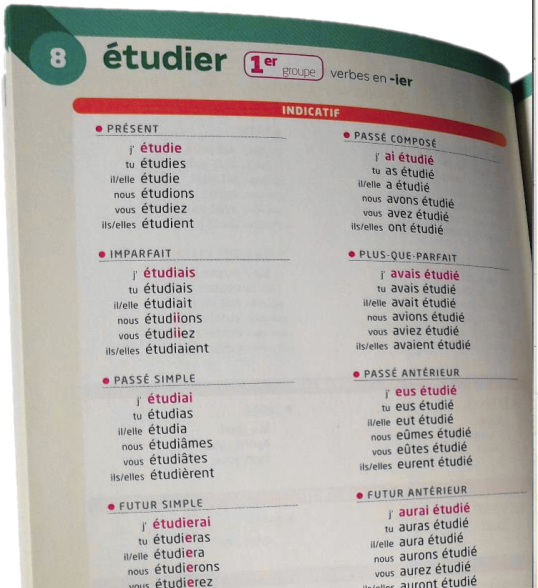
PASSÉ
avoir étudié

PARTICIPE

PRÉSENT

PASSÉ
ayant étudié
étudié(e, s, es)

Inflection



8 étudier 1er groupe verbes en -ier

INDICATIF

• PRÉSENT
j' étudie
tu étudies
il/elle étudie
nous étudions
vous étudiez
ils/elles étudient

• PASSÉ COMPOSÉ
j' ai étudié
tu as étudié
il/elle a étudié
nous avons étudié
vous avez étudié
ils/elles ont étudié

• IMPARFAIT
j' étudiais
tu étudiais
il/elle étudiait
nous étudiions
vous étudiiez
ils/elles étudiaient

• PLUS-QUE-PARFAIT
j' avais étudié
tu avais étudié
il/elle avait étudié
nous avions étudié
vous aviez étudié
ils/elles avaient étudié

• PASSÉ SIMPLE
j' étudiai
tu étudias
il/elle étudia
nous étudiâmes
vous étudiâtes
ils/elles étudièrent

• PASSÉ ANTÉRIEUR
j' eus étudié
tu eus étudié
il/elle eut étudié
nous eûmes étudié
vous eûtes étudié
ils/elles eurent étudié

• FUTUR SIMPLE
j' étudierai
tu étudieras
il/elle étudiera
nous étudierons
vous étudierez

• FUTUR ANTÉRIEUR
j' aurai étudié
tu auras étudié
il/elle aura étudié
nous aurons étudié
vous aurez étudié

Ire DÉCLINAISON DES NOMS

GEN. SG. : -A E

(FÉMININS — quelques masculins)

12. Type : **rosa, ae**, f. : *la rose, une rose.*

			Terminaisons
SINGULIER	NOMINATIF	rosa	-a
	VOCATIF	rosa	-a
	ACCUSATIF	rosam	-am
	GÉNITIF	rosae	-ae
	DATIF	rosae	-ae
	ABLATIF	rosā	-ā
PLURIEL	NOMINATIF	rosae	-ae
	VOCATIF	rosae	-ae
	ACCUSATIF	rosas	-as
	GÉNITIF	rosárum	-árum
	DATIF	rosis	-is
	ABLATIF	rosis	-is

PARTICULARITÉS DE LA Ire DÉCLINAISON

13. **Família**, ae, f. : *famille* a gardé un génitif sg. archaïque en **-as** dans paterfamílias, patrisfamílias, m. *père de famille* — materfamílias, matrisfamílias, f. *mère de famille*

14. **Dea**, deae, f. : *déesse* et **fília**, ae, f. : *fille* font au datif et à l'ablatif pluriels **deábus** et **filiábus** quand il faut éviter une confusion avec le dat.-abl. pl. de deus, dei, m. : *dieu* et fílius, ii, m. : *fils*.

Deis et **deábus** : *Aux dieux et* ***aux déesses.***
Fíliis et **filiábus** : *Aux fils et* ***aux filles.***

➔ Mais on peut trouver **deis, fíliis** venant de **dea, fília**, s'il n'y a pas de risque de confusion.

15. Formes poétiques : gén. sg. en **-ái** au lieu de **-ae** : **terrái**
gén. pl. en **-um** au lieu de **-árum** : **agricolum.**

This talk

1. Why am I interested in data management ?
2. Current problems
3. Data principles and concrete step to apply them
4. The Paralex standard for inflected lexicons

Why?

Variability in paradigm sizes

Vietnamese: 1 form

(1) Tôi học
I study
'I study'

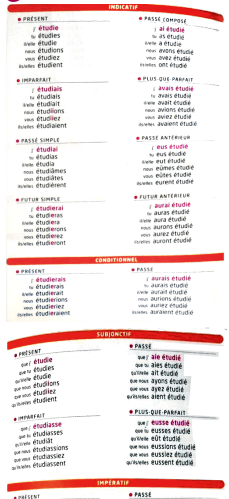
(2) Tôi đang học
I FUT study
'I will study'

(3) Tôi đã học
I PST study
'I studied'

English: 5 forms

	STUDY
PRS	stʌdɪ
PRS.3SG	stʌdɪz
PST	stʌdɪd
PST.PART	stʌdɪd
PRS.PART	stʌdɪɪŋ

French: 50 forms



Archi: 1 500 000 forms (Dagestan)

Variability in paradigm structures

English

cat	‘one cat’
cat-s	‘many cats’

Walpiri (Australia)

wati	‘one man’
wati-jarra	‘two men’
wati-patu	‘several men’

Yei (New Guinea)

yə-mdəd-ə	‘one sat’
yɛ-mdəd-anɛ	‘two sat’
yɛ-mdəd-ə	‘many sat’

Nuer (Soudan)

ṯṳ̈rbîl	‘one cat’
ṯṳ̈rbîɛɛl	‘of a cat’
ṯṳ̈̽rbĭɛɛlḭ́	‘many cats’

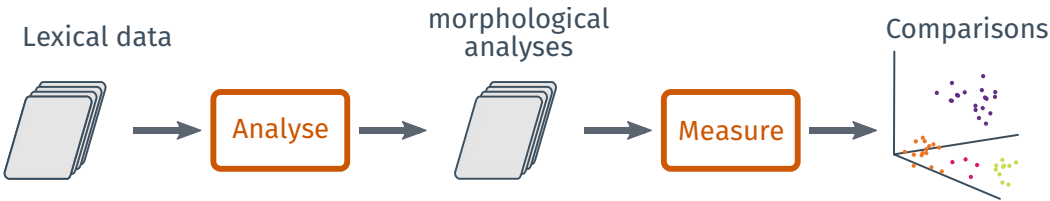
Research questions

(1) *What is possible?*
characterize the variability space

(2) *What is observed ?*
Find macro-linguistic variation distributions

(3) *Why?*
Testable hypotheses on cognition, cultural evolution, communication

Computational methodology

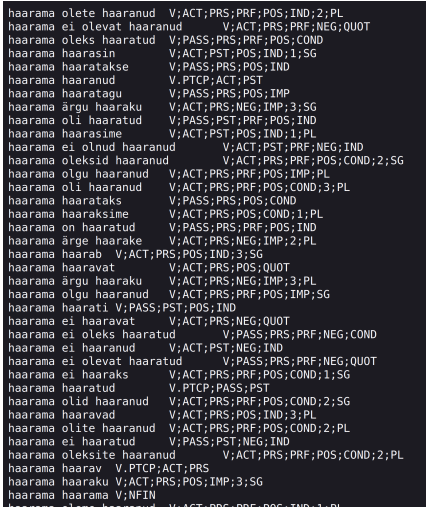


Large scale & high resolution multivariate typology:

- Precise, high quality data;
- Automated linguistic analyses
- Quantitative measurements

Why not the unimorph datasets ?

Kirov et al. (2016) and Batsuren et al. (2022):



```
haarama olete haaranud  V;ACT;PRS;PRF;POS;IND;2;PL
haarama ei olevat haaranud      V;ACT;PRS;PRF;NEG;QUOT
haarama oleks haaratud  V;PASS;PRS;PRF;POS;COND
haarama haarasin        V;ACT;PST;POS;IND;1;SG
haarama haaratakse      V;PASS;PRS;POS;IND
haarama haaranud        V.PTCP;ACT;PST
haarama haaratagu       V;PASS;PRS;POS;IMP
haarama ärgu haaraku    V;ACT;PRS;NEG;IMP;3;SG
haarama oli haaratud    V;PASS;PST;PRF;POS;IND
haarama haarasime       V;ACT;PST;POS;IND;1;PL
haarama ei olnud haaranud       V;ACT;PST;PRF;NEG;IND
haarama oleksid haaranud        V;ACT;PRS;PRF;POS;COND;2;SG
haarama olgu haaranud   V;ACT;PRS;PRF;POS;IMP;PL
haarama oli haaranud    V;ACT;PRS;PRF;POS;COND;3;PL
haarama haarataks       V;PASS;PRS;POS;COND
haarama haaraksime      V;ACT;PRS;POS;COND;1;PL
haarama on haaratud     V;PASS;PRS;PRF;POS;IND
haarama ärge haarake    V;ACT;PRS;NEG;IMP;2;PL
haarama haarab  V;ACT;PRS;POS;IND;3;SG
haarama haaravat        V;ACT;PRS;POS;QUOT
haarama ärgu haaraku    V;ACT;PRS;NEG;IMP;3;PL
haarama olgu haaranud   V;ACT;PRS;PRF;POS;IMP;SG
haarama haarati V;PASS;PST;POS;IND
haarama ei haaravat     V;ACT;PRS;NEG;QUOT
haarama ei oleks haaratud       V;PASS;PRS;PRF;NEG;COND
haarama ei haaranud     V;ACT;PST;NEG;IND
haarama ei olevat haaratud      V;PASS;PRS;PRF;NEG;QUOT
haarama ei haaraks      V;ACT;PRS;PRF;POS;COND;1;SG
haarama haaratud        V.PTCP;PASS;PST
haarama olid haaranud   V;ACT;PRS;PRF;POS;COND;2;SG
haarama haaravad        V;ACT;PRS;POS;IND;3;PL
haarama olite haaranud  V;ACT;PRS;PRF;POS;COND;2;PL
haarama ei haaratud     V;PASS;PST;NEG;IND
haarama oleksite haaranud       V;ACT;PRS;PRF;POS;COND;2;PL
haarama haarav  V.PTCP;ACT;PRS
haarama haaraku V;ACT;PRS;POS;IMP;3;SG
haarama haarama V;NFIN
```

- **Crowdsourced** origin
- Minimalist, **canonical** structure
- Strictly **orthographic**
- Large but **noisy**

https://unimorph.github.io/

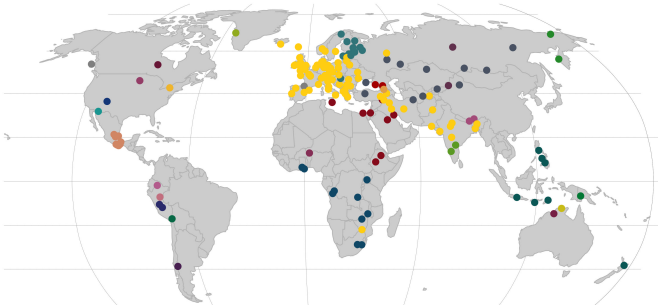
Problems

Culture

"You can also collect butterflies and make many observations. If you like butterflies, that's fine; but such work must not be confounded with research, which is concerned to discover explanatory principles of some depth and fails if it does not do so."

Chomsky and Ronat, 1979, p. 57

Coverage

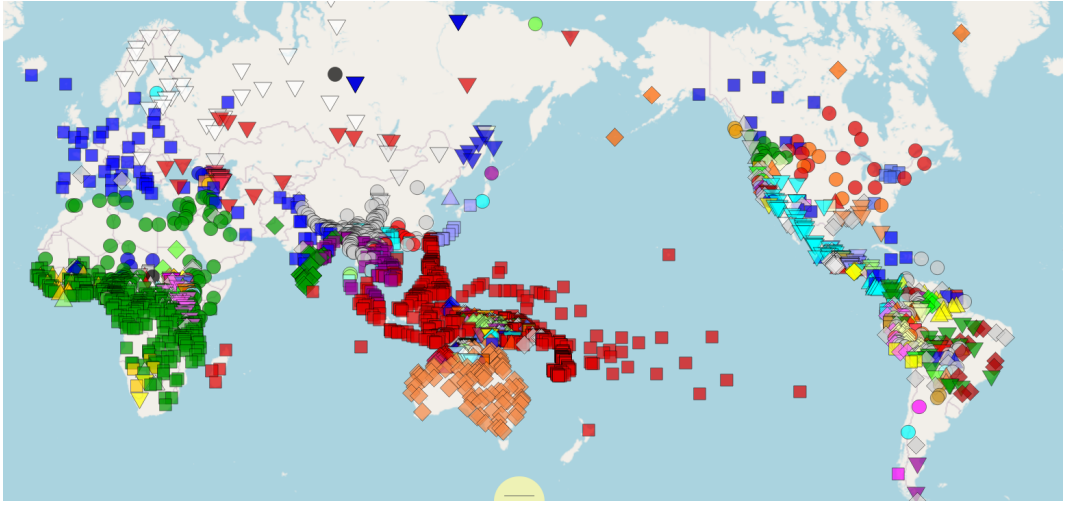


Macro-family

- Abkhaz-Adyge
- Afro-Asiatic
- Algic
- Araucanian
- Arawakan
- Athabaskan-Eyak-Tlingit
- Atlantic-Congo
- Austronesian
- Aymaran
- Bosavi
- Chukotko-Kamchatkan
- Dravidian
- Eskimo-Aleut
- Gunwinyguan
- Indo-European
- Iroquoian
- Kartvelian
- Otomanguean
- Pano-Tacanan
- Quechuan
- Sino-Tibetan
- Siouan
- Songhay
- Southern Daly
- Tungusic
- Turkic
- Uralic
- Uto-Aztecan
- NA

Coverage

Grambank (Skirgård et al., 2023): https://grambank.clld.org/



Consistency and machine-readability

Gap between human- and machine-readability:

- **Formats**
- **Coding**
- Down to details of encoding

Commensurability

- Linguistic data are analyses
- Information of absence vs absence of information
- Interpretation of coding schemes

Durability

Romance Verbal Inflection Dataset 2.0 Beniamine et al (2020)

Data imprisoned within its own website

web scraping, cleaning, normalisation, standardisation, publication

https://doi.org/10.5281/zenodo.4039059

Surrey Morphological Complexity Database

Data initially published in Flash, unsupported since 2021.

Data archeology, cleaning, updating, normalisation, publication

http://dx.doi.org/10.15126/SMG.23/1

Technical skills

Rarely do we teach linguistics students about:

- Machine readable formats (xml, json, csv)
- Versioning (git)
- Data cleaning, validation, testing
- Metadata, standards

Some solutions

Data principles

Some ⌨ Paralex collaborators

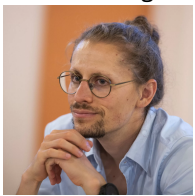
Cormac Anderson

Jules Bouton

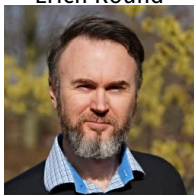
Mae Carroll

Borja Herce

Matteo Pellegrini



Erich Round



Helen Sims-Williams

DeAR Principles

- **Decentralized**
  - Centralized data is neither robust nor long-lasting
  - International collaboration must be incentivised
- **Automatically validated**
  - Manual curation of large datasets is necessary but error-prone
  - Automatic quality control of structure and content is crucial
- **Revisable**
  - Data evolves, its presentation must be updateable
  - Seamless updates of showcases can be generated from a single data source

Beniamine et al 2023. *Paralex: a DeAR standard for rich lexicons of inflected forms*. ISMo: International Symposium of Morphology.

Solutions

To publish high quality, easily citable, scientifically impactful data, useful for the long term:

Creation

- Versioning DeAR
- Metadata FAIR DeAR
- Standards Inter-operable Reusable DeAR
- Linked data Interoperable Reusable
- Validation DeAR

Publication

- Documentation Reusable
- Licenses Open Reusable
- Archived downloads FAIR
- DOIs Findable Accessible
- Continuous pipelines DeAR

Standards

For data points

- Metric systems
- Leipzig glossing rules
- ISO 639 (langues)
- ISO 3166 (pays)

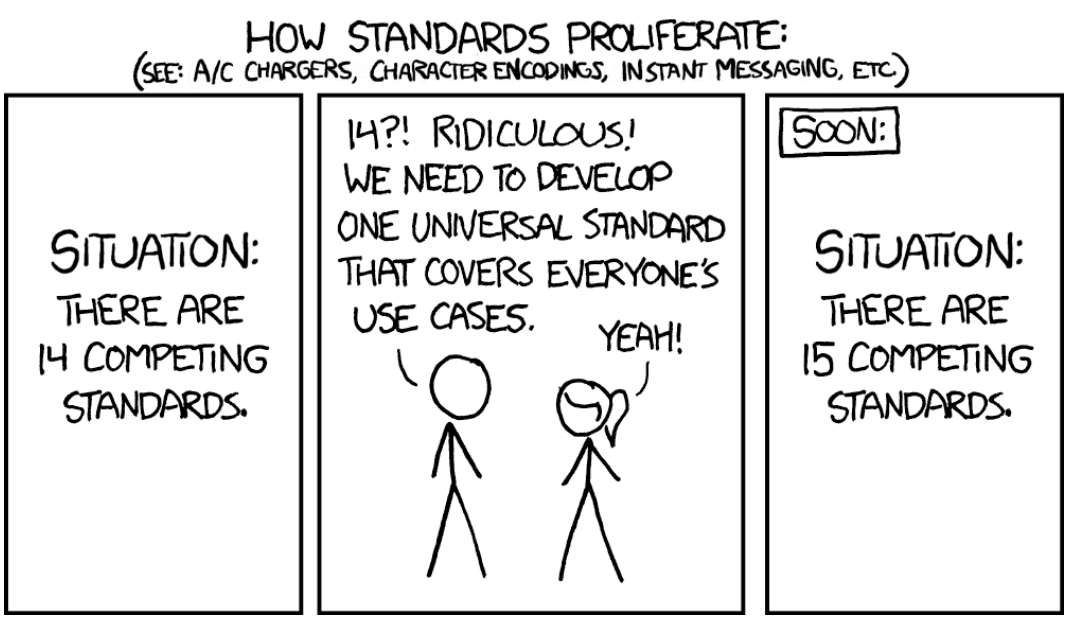
For datasets

- Text Encoding Initiative (TEI, XML)
- CLDF
- CONLL-U
- Paralex !

For metadata

- Frictionless (json)
- Dublin Core (xml)

The Paralex standard


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION:
THERE ARE
14 COMPETING
STANDARDS.
14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE'S
USE CASES.
YEAH!
SOON:
SITUATION:
THERE ARE
15 COMPETING
STANDARDS.

Paralex

http://www.paralex-standard.org

Paralex

sbeniamine/paralex
☆0 ⑂1

Paralex
Home
Standard
Specs
Ontology
Background
Tutorial
Datasets
Changelog

Paralex: lexicons of morphological paradigms

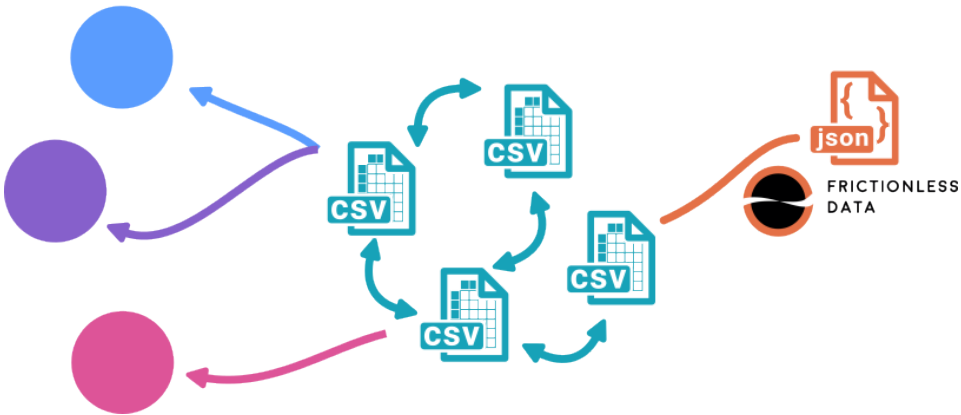
pypi package 2.2.1

Paralex is a standard for morphological lexicons which document inflectional paradigms.

It strives to provide data which is FAIR, so it can be used automatically, CARE, so it respects and empowers language communities, and DeAR (our own set of principles), so we can create a virtuous data ecosystem. It was inspired by the Cross-Linguistic Data Formats (CLDF) standard, and adheres to a similar philosophy. We aim to keep the two standards compatible in order to facilitate inter-operability.

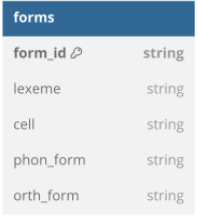
Paralex

LINKED DATA METADATA



Relational schema

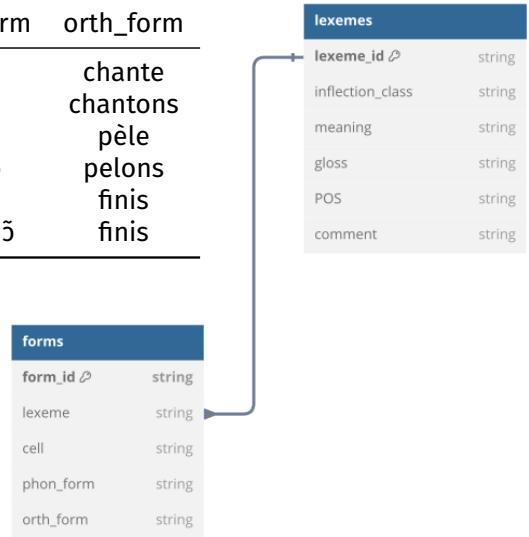
form_id	lexeme	cell	phon_form	orth_form
f1	CHANTER	PRS.IND.1.SG	ʃ ɑ̃ t	chante
f4	CHANTER	PRS.IND.1.PL	ʃ ɑ̃ t ɔ̃	chantons
f60	PELER	PRS.IND.1.SG	p ɛ l	pèle
f64	PELER	PRS.IND.1.PL	p ø l ɔ̃	pelons
f90	FINIR	PRS.IND.1.SG	f i n i	finis
f94	FINIR	PRS.IND.1.PL	f i n i s ɔ̃	finis



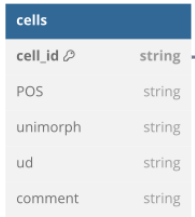
Relational schema

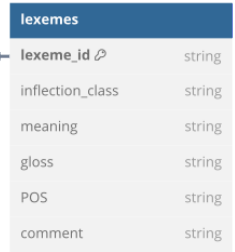
form_id	lexeme	cell	phon_form	orth_form
f1	CHANTER	PRS.IND.1.SG	ʃ ɑ̃ t	chante
f4	CHANTER	PRS.IND.1.PL	ʃ ɑ̃ t ɔ̃	chantons
f60	PELER	PRS.IND.1.SG	p ɛ l	pèle
f64	PELER	PRS.IND.1.PL	p ø l ɔ̃	pelons
f90	FINIR	PRS.IND.1.SG	f i n i	finis
f94	FINIR	PRS.IND.1.PL	f i n i s ɔ̃	finis

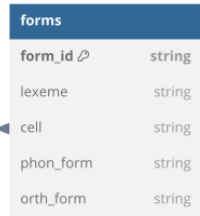
lexeme_id	inflection_class	gloss
CHANTER	groupe-1	to eat
PELER	groupe-1	to peel
FINIR	groupe-2	to end



Relational schema







cell_id	unimorph	POS
IND.PRS.1.SG	V;IND;PRS;1;SG	verb
IND.PRS.1.PL	V;IND;PRS;1;PL	verb

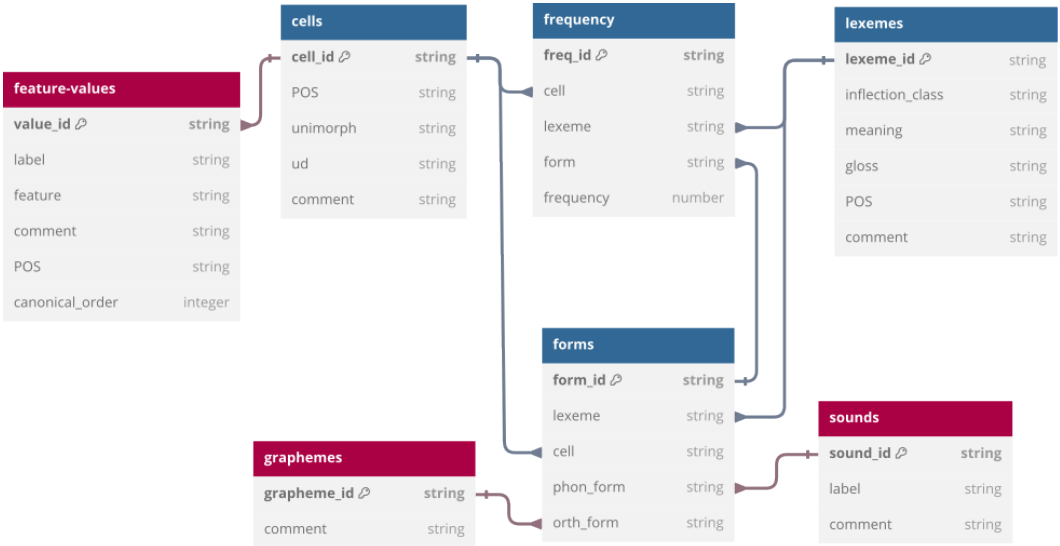
Relational schema

cell_id	GRACE	flexique	unimorph	ud	ftb
COND.PRS.1.PL	Vmcp1p-	cond.1pl	V;COND;1;PL	Mood=Cnd|Number=Plur|Person=1|Tense=Pres	m=cond|n=p|p=1|t=pst
COND.PRS.1.SG	Vmcp1s-	cond.1sg	V;COND;1;SG	Mood=Cnd|Number=Sing|Person=1|Tense=Pres	m=cond|n=s|p=1|t=pst
COND.PRS.2.PL	Vmcp2p-	cond.2pl	V;COND;2;PL	Mood=Cnd|Number=Plur|Person=2|Tense=Pres	m=cond|n=p|p=2|t=pst
COND.PRS.2.SG	Vmcp2s-	cond.2sg	V;COND;2;SG	Mood=Cnd|Number=Sing|Person=2|Tense=Pres	m=cond|n=s|p=2|t=pst
COND.PRS.3.PL	Vmcp3p-	cond.3pl	V;COND;3;PL	Mood=Cnd|Number=Plur|Person=3|Tense=Pres	m=cond|n=p|p=3|t=pst
COND.PRS.3.SG	Vmcp3s-	cond.3sg	V;COND;3;SG	Mood=Cnd|Number=Sing|Person=3|Tense=Pres	m=cond|n=s|p=3|t=pst
IND.FUT.1.PL	Vmif1p-	fut.1pl	V;IND;FUT;1;PL	Mood=Ind|Number=Plur|Person=1|Tense=Fut	m=ind|n=p|p=1|t=fut
IND.FUT.1.SG	Vmif1s-	fut.1sg	V;IND;FUT;1;SG	Mood=Ind|Number=Sing|Person=1|Tense=Fut	m=ind|n=s|p=1|t=fut
IND.FUT.2.PL	Vmif2p-	fut.2pl	V;IND;FUT;2;PL	Mood=Ind|Number=Plur|Person=2|Tense=Fut	m=ind|n=p|p=2|t=fut
IND.FUT.2.SG	Vmif2s-	fut.2sg	V;IND;FUT;2;SG	Mood=Ind|Number=Sing|Person=2|Tense=Fut	m=ind|n=s|p=2|t=fut
IND.FUT.3.PL	Vmif3p-	fut.3pl	V;IND;FUT;3;PL	Mood=Ind|Number=Plur|Person=3|Tense=Fut	m=ind|n=p|p=3|t=fut
IND.FUT.3.SG	Vmif3s-	fut.3sg	V;IND;FUT;3;SG	Mood=Ind|Number=Sing|Person=3|Tense=Fut	m=ind|n=s|p=3|t=fut
IMP.PRS.1.PL	Vmmp1p-	imp.1pl	V;POS;IMP;1;PL	Mood=Imp|Number=Plur|Person=1|Tense=Pres	m=Imp|n=p|p=1|t=pst
IMP.PRS.2.PL	Vmmp2p-	imp.2pl	V;POS;IMP;2;PL	Mood=Imp|Number=Plur|Person=2|Tense=Pres	m=Imp|n=p|p=2|t=pst
IMP.PRS.2.SG	Vmmp2s-	imp.2sg	V;POS;IMP;2;SG	Mood=Imp|Number=Sing|Person=2|Tense=Pres	m=Imp|n=s|p=2|t=pst
INF	Vmn—-	inf	V;NFIN	VerbForm=Inf	m=inf
IND.IPFV.1.PL	Vmii1p-	ipfv.1pl	V;IND;PST;1;PL;IPFV	Mood=Ind|Number=Plur|Person=1|Tense=Imp	m=ind|n=p|p=1|t=Imp
IND.IPFV.1.SG	Vmii1s-	ipfv.1sg	V;IND;PST;1;SG;IPFV	Mood=Ind|Number=Sing|Person=1|Tense=Imp	m=ind|n=s|p=1|t=Imp
IND.IPFV.2.PL	Vmii2p-	ipfv.2pl	V;IND;PST;2;PL;IPFV	Mood=Ind|Number=Plur|Person=2|Tense=Imp	m=ind|n=p|p=2|t=Imp
IND.IPFV.2.SG	Vmii2s-	ipfv.2sg	V;IND;PST;2;SG;IPFV	Mood=Ind|Number=Sing|Person=2|Tense=Imp	m=ind|n=s|p=2|t=Imp
…	…	…	…	…	

Vocabulary relations

form_id	lexeme	cell	phon_form
form_158742	mésuser	ptcp.pst.f.pl	m E z y z e
form_819	aboyer	ind.prs.1.pl	a b w a j ɔ̃
form_41745	chroniquer	cond.prs.1.sg	k ʁ O n i k ə ʁ E
form_91334	détricoter	ind.pst.3.pl	d E t ʁ i k O t E ʁ
form_197935	refleurir	imp.prs.2.sg	ʁ ə f l Ø ʁ i
form_122951	galvaniser	ind.pst.3.sg	g a l v a n i z a
form_11785	anoblir	ind.prs.2.pl	a n O b l i s e
form_99328	encourager	ind.prs.2.pl	ɑ̃ k u ʁ a ʒ e
form_237143	surprendre	sbjv.prs.1.pl	s y ʁ p ʁ ə n j ɔ̃
…	…	…	…

Vocabulary relations



Frequencies

We need frequencies to go beyond idealized paradigm tables:

- Effect on paradigm structure
- Studying non-canonical phenomena

But obtaining good frequencies is hard:

A Skewed frequencies in annotated corpora

B Syncretic forms within lexemes

C Homonymous forms across lexemes

Orthographic tokens and inflected forms

(4) a. Elle savait que vous **veniez**
she.3SG.F knew.IMP.3SG COMP you.2PL come.**IND.IPFV.2PL**
'She knew that you were coming.'

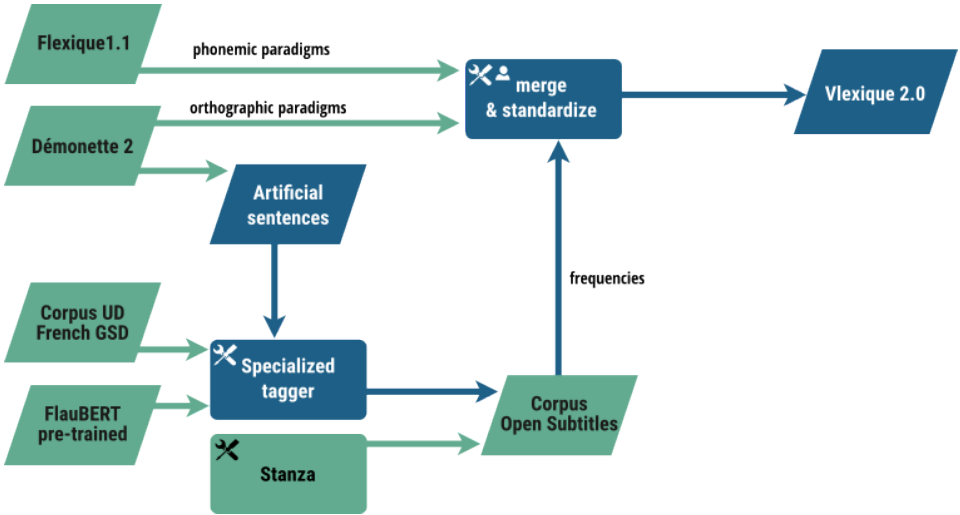
b. Il faudrait que vous **veniez**
one.3SG.M must.COND.3SG COMP you.2PL come.**SBJV.PRS.2PL**
'You should come.'

(5) a. Je **comparais** devant une cour [...] (**COMPARAÎTRE**)
I.1SG appear.IND.PRS.1.SG before a court
'I appear before a court'

b. Je **comparais** juste les tailles ! (**COMPARER**)
I.1SG compare.IND.IPFV.1.SG only the sizes !
'I was only comparing sizes' !

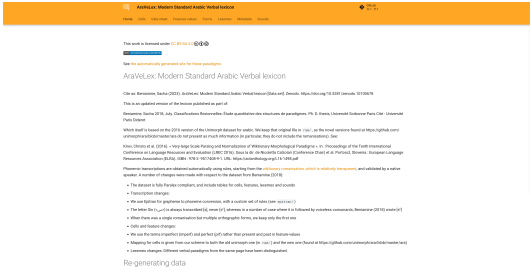
From Open Subtitles (Lison and Tiedemann, 2016)

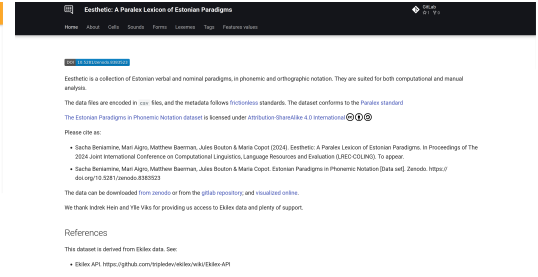
Vlexique creation

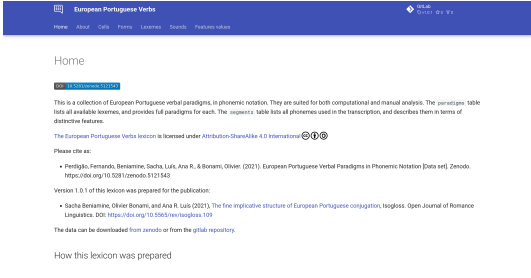


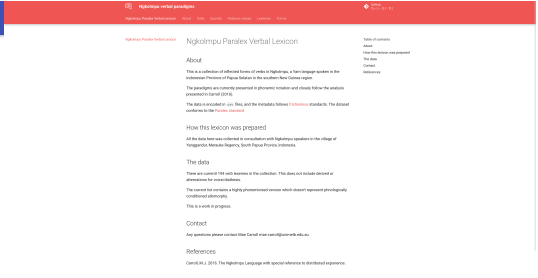
Beniamine, Coavoux & Bonami (2024). *Vlexique2.0: A rich lexicon of French verbal inflection with form-level frequencies*. 21st International Morphology Meeting, Aug 2024, Vienna, Austria. hal-04689352

Generated websites

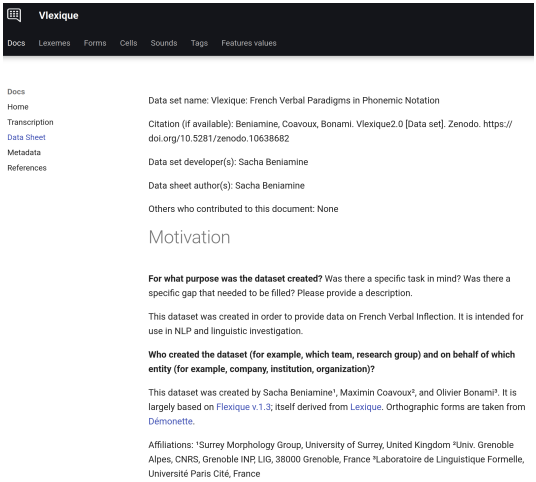








Documentation and data sheet



- Document analytical choices
- Document any hesitations or decisions in coding
- Data sheet: targeted questions mixing data statements (Bender and Friedman, 2018) & data sheets (Gebru et al., 2021)

Current coverage

Yam (more1255) | Pama-Nyungan (pama1250) | Uralic (ural1272) | Indo-European (indo1319)

Otomanguean (otom1299) | Atlantic-Congo (atla1278) | Sino-Tibetan (sino1245) | Afro-Asiatic (afro1255)

Conclusion

- Computational linguistics benefits from datasets which are:
  - high quality
  - long lasting
  - open, FAIR, CARE, & DeAR
- Large noisy datasets are not enough
- A crucial point is to lower the technical bar
- Good data are among the most important contributions we can make to linguistics

Conclusion

- Computational linguistics benefits from datasets which are:
  - high quality
  - long lasting
  - open, FAIR, CARE, & DeAR
- Large noisy datasets are not enough
- A crucial point is to lower the technical bar
- Good data are among the most important contributions we can make to linguistics

Thanks !